Evaluating Adversarial Partitions

Andreas Pashalidis and Stefan Schiffner

K.U. Leuven/IBBT, ESAT/SCD-COSIC
Kasteelpark Arenberg 10,
Leuven, Belgium
{andreas.pashalidis,stefan.schiffner}@esat.kuleuven.be

Abstract. In this paper, we introduce a framework for measuring unlinkability both per subject and for an entire system. The framework enables the evaluator to attach different sensitivities to individual items in the system, and to specify the severity of different types of error that an adversary can make. These parameters, as well as a threshold that defines what constitutes a privacy breach, may be varied for each subject in the system; the framework respects and combines these potentially differing parametrisations. It also makes use of graphs in a way that results in *intuitive* feedback of different levels of detail. We exhibit the behaviour of our measures in two experimental settings, namely that of adversaries that output randomly chosen partitions, and that of adversaries that launch attacks of different effectiveness.

1 Introduction

An adversary on unlinkability aims to divide a given set of elements into non-overlapping clusters, such that the elements in each cluster belong to the same subject or, more generally, share a well-defined property. In the electronic world, such elements are typically digital data items that arise as a result of online transactions, e.g. personal messages, shopping records, user attributes, protocol transcripts, or entries in an audit log. Measuring linkability is important because an adversary's ability to link elements that should be unlinkable constitutes a privacy breach. Moreover, successful attacks on linkability can lead to further privacy breaches such as the unauthorised (re-)identification of subjects [13]. We stress, however, that an adversary on unlinkability does not necessarily care about identifying subjects.

Linkability measurements are not straightforward. Consider, for example, the partition $\Pi = \{\{\circ, \circ, \circ, \circ, \circ\}, \{\bullet, \bullet\}\}$, and let us call its two clusters 'Alice's and Bob's items'. An adversary assuming that the correct partition is $\{\{\circ, \circ, \circ, \circ, \circ\}, \{\bullet\}, \{\bullet\}\}$, obviously failed to link Bob's items, but has linked Alice's items perfectly. It is, however, not obvious which of

the partitions, $\Pi'_1 = \{\{\circ, \circ, \circ, \circ\}, \{\circ, \bullet, \bullet\}\}$, $\Pi'_2 = \{\{\circ, \circ, \circ, \bullet\}, \{\circ, \circ\}, \{\bullet\}\}$, and $\Pi'_3 = \{\{\circ, \circ, \circ\}, \{\circ\}, \{\circ\}, \{\bullet, \bullet\}\}$, for example, is a better approximation of Π, both overall and with respect to any given subject. In order to demonstrate this, we first focus on Alice's items but ignore, for the moment, Bob's. According to Π'_1, almost all her items have been identified as belonging together, with the exception of one, which belongs to a different cluster. Π'_2 also divides Alice's items into two clusters, but the second one contains two items rather than only one. Intuition therefore suggests that Π'_2 is a worse approximation than Π'_1, at least with respect to Alice's items. Π'_3 divides Alice's items into three clusters. However, one has to re-allocate the same amount of items (two) in both Π'_2 and Π'_3 to completely link Alice's items. This, however, does not necessarily mean that both clusters are equally good approximations; Π'_3 may still be considered worse than Π'_2 since, in Π'_2, a single cluster *merging* suffices to completely link Alice's items, while, in Π'_3, two such mergings are required. We now stop ignoring Bob's items, but still focus on unlinkability from Alice's point of view. While Π'_1 contains a cluster that mixes two of Bob's items with some of Alice's, Π'_2 mixes only one. Moreover, Π'_3 contains no cluster that mixes items of both Alice and Bob. This second viewpoint leads to opposite conclusions: Π'_3 is a better approximation of Π than Π'_2, and Π'_2 is better than Π'_1.

In order to decide which of the two viewpoints should prevail, it is important to know more about the concrete context. A loan seeker, for example, would like to prevent his bank from being able to link his negative credit ratings, but if linking happens anyway, then he is likely to prefer correct rather than incorrect inferences to be made.[1] A consumer, on the other hand, would like to prevent the direct marketing company from being able to aggregate his shopping behaviour into a detailed profile, but if this happens anyway, he is likely to prefer his profile to be 'contaminated' with the shopping histories of other people. Both the loan seeker and the consumer are therefore interested in both viewpoints, but have different preferred outcomes with respect to the latter viewpoint.

In this paper, we develop a novel evaluation framework for adversarial partitions that enables one to take these different viewpoints. The framework is flexible in multiple ways, as it enables the evaluator to attach different sensitivities to individual items in the system, to specify the severity of different types of error that an adversary can make, and to define a threshold that defines what constitutes a privacy breach. These

[1] We assume that the loan seeker is willing to pay back his own debts in order to erase negative credit ratings if necessary, but unwilling to do so for other people's debts.

parametrisations may vary for each subject in the system; the framework respects and combines them. Evaluations are communicated by means of intuitive graphs and, since they are performed on the subject level, can serve as the basis for further analysis such as fairness.

The rest of this paper is organised as follows. The next section surveys related work, and section 3 introduces our evaluation framework. Section 4 illustrates its application on a toy example. Section 5 compares the behaviour of our measures to that of other distance measures. Finally, section 6 concludes.

2 Related work

Two research areas relate to our work, namely that of privacy and that of statistical classification. In particular, we build on ideas on measuring unlinkability from the former, and comparing clusterings from the latter area. Works from the first area deal with the question of how effectively certain privacy preserving systems protect the unlinkability of the elements that arise in the system. While [3, 4, 9, 11], for example, measure unlinkability in general, [13, 14] and [1] do so in the context of anonymous communication and attribute sharing, respectively. In addition, [11] considers fairness.

The literature on comparing clusterings, on the other hand, has a longer history, and many distance metrics on partitions have been defined and used over the years [5, 15]. The Rand index [12], for example, considers the extent to which two partitions treat all element pairs similarly, the minimum transfer distance [2] considers the number of element *transfers* until two partitions are identical and the variation of information [7] uses information-theoretic primitives.

The overall goal of both research areas is to provide the basis to compare partitions. The difference is that, in the 'measuring unlinkability' area, these partitions represent attacks on specific privacy preserving systems, while, in the 'comparing clusterings' context, they represent algorithms that aim to classify the items of a given dataset in some useful way. Unfortunately, the approaches from both areas suffer from certain inflexibilities that limit their suitability when it comes to evaluating adversarial partitions in a privacy setting. These inflexibilities manifest themselves in three different ways, as follows.

Firstly, existing approaches do not take into account important aspects of adversarial partitions. While [14], for example, mainly focuses on the question of whether or not *two* given elements are linked, [3, 9] as

well as the literature on comparing clusterings, compute measurements over a given partition as a whole. Both approaches do not measure unlinkability on the crucial *subject* level. Other works, e.g. [4, 11, 14], also consider the (degree of) unlinkability of arbitrary element subsets; their focus is, however, on the extent to which the elements of the subset can be linked, while the extent to which foreign elements (i.e. elements not in the subset) 'contaminate' the adversary's view on the subset, are disregarded. (Here it should be mentioned that this is not entirely true for the 'white-box' analysis approach described in [3]; depending on the chosen partition distance metric, contamination levels may be taken into account. This, however, does not happen on the subject level, but rather on the overall partition level.) In this work, we ask the question 'how well is a given *solution cluster* hidden within the adversarial partition?'. That is, we perform separate measurements pertaining to each subject without ignoring foreign elements (i.e. elements belonging to other subjects), and, if necessary, combine these into an overall average only in the last step. This approach yields not only measurements on the subject level, but also natural ways to evaluate the *fairness* provided by the underlying system.

Secondly, existing approaches do not distinguish between the sensitivity that users or evaluators attach to the elements in a system. They also do not let the evaluator specify his sensitivity towards different types of error that an adversarial partition may exhibit. That is, existing approaches do not account for the fact that the same adversarial partition may represent attacks of different seriousness in different contexts. Our evaluation framework enables the evaluator to formulate such sensitivities.

Thirdly, it is unclear how to construct supportive material, such as illustrations or graphics, that show, in a sufficiently intuitive way, how a given (un)linkability measurement comes about. While the graphs that represent an attacker's internal state as defined in [9] are certainly an exception, they do not convey the information of how well a given subject's cluster is hidden within the adversary's state (especially in the presence of 'transitivity contradictions'). Our evaluation framework uses graphs that depict meaningful quantities that individual subjects are likely to care about. In fact, in our framework, these graphs do not depict the final unlinkability measurements. Rather conversely, the final unlinkability measurements are *derived* from the graphs.

Our approach remains agnostic to specific applications, and combines ideas from both research areas above. From the literature of measuring unlinkability we follow the idea that unlinkability generally decreases as

the adversary links more of a given subject's elements, and we take into account adversaries that output multiple, in their view probable partitions. From the literature on comparing clusterings, we adopt some (very) basic notions of [7]. Namely, the intersection of cluster pairs is an important quantity that defines, among other parameters, both 'miss' and 'include' error counts in our approach. These same intersections also play a central role in the 'variation of information' metric (see Equation 15 in [7]). However, it is not our goal to define distance metrics in the strict sense; therefore our framework does not aim to fulfill the axioms put forth in [8]. It does, however, fulfill the informal criteria listed in [3], namely taking into account both the certainty and the consistency of the adversary.

3 Evaluating adversarial partitions

Our evaluation methodology focuses on the errors made by an adversary, and distinguishes between primary and secondary errors. The motivation for this distinction lies in the 'sort of story' that the two error types tell: primary errors describe the adversary's current state, while secondary errors describe the risk that this state represents for the future. More precisely, primary errors describe how well a given solution cluster is *currently* hidden within a given adversarial partition, while secondary errors describe how well the cluster *remains* hidden if additional information would enable the adversary to further refine its current assessment. In the following, Π denotes the solution partition, and Π' the adversarial partition. We assume that both Π and Π' are set partitions of the *same* finite set of size $n = \sum_{C \in \Pi} |C|$ and use the notation C and C' to refer to individual clusters of Π and Π', respectively.

3.1 Primary errors

Motivated by the example in the introduction, we consider two types of primary error that an adversary can make with respect to a solution cluster: a 'miss' error occurs if the adversary fails to include an element that should be included in the cluster, and an 'include' error occurs if it includes an element that should not have been included. Formally, given a cluster pair C, C', the number of miss and include errors is defined as $m(C, C') = |C - C'|$ and $i(C, C') = |C' - C|$, respectively. The miss (resp. include) error counts can also be defined as $m(C, C') = |C| - |C \cap C'|$ (resp. $i(C, C') = |C'| - |C \cap C'|$). This makes explicit the intersection mentioned in section 2.

The evaluator may be more sensitive towards miss than he is towards include errors, or vice versa. We let the evaluator indicate this sensitivity by means of a 'policy parameter' $\alpha \in [0,1]$: setting $\alpha = 0$ indicates that he cares exclusively about the extent to which the elements of interest are linked, while completely ignoring foreign elements. Setting $\alpha = 1$, on the other hand, means that the evaluator is agnostic to the extent to which the adversary has managed to link the elements, and is only interested in the extent to which foreign elements are being mixed together with the correct ones. If both error types are to be deemed equally important, then the evaluator has to set $\alpha = 1/2$.

We are interested to count errors per subject, i.e. *per solution cluster*. One approach would be, given a solution cluster, to average the error counts over all adversarial clusters. However, since only few adversarial clusters are likely to be related to any given solution cluster C, we do not follow this approach. Instead we use only the *most relevant* adversarial cluster as the basis to count primary errors. This cluster is one of the adversarial clusters that have at least one element in common with C, and we use the policy parameter α to identify which one it is. We proceed as follows. Given C, we first determine the subset of related clusters $L(C, \Pi') = \{C' \in \Pi' : |C \cap C'| \geq 1\}$. Then, we calculate the *priority* for each cluster in $L(C, \Pi')$ as $p_\alpha(C, C') = (\alpha m(C, C') + (1-\alpha) i(C, C'))^{-1}$. The most relevant adversarial cluster is the one with the highest priority.

If two or more related adversarial clusters yield this maximum, then we take the average of their error counts. The complete formal procedure is therefore as follows. Given a solution cluster C, we first determine the set of most relevant adversarial clusters as $R_\alpha(C, \Pi') = \{C' \in L(C, \Pi') : p_\alpha(C, C') = \max_{C' \in L(C,\Pi')} p_\alpha(C, C')\}$. Then the error counts for solution cluster C with respect to an adversarial partition Π' are given by $m(C, \Pi', \alpha) = \sum_{C' \in R_\alpha(C,\Pi')} {}^{m(C,C')}/_{|R_\alpha(C,\Pi')|}$ and $i(C, \Pi', \alpha) = \sum_{C' \in R_\alpha(C,\Pi')} {}^{i(C,C')}/_{|R_\alpha(C,\Pi')|}$. Finally, the total error counts are obtained simply by adding up the per solution cluster counts: $m(\Pi, \Pi', \alpha) = \sum_{C \in \Pi} m(C, \Pi', \alpha)$ and $i(\Pi, \Pi', \alpha) = \sum_{C \in \Pi} i(C, \Pi', \alpha)$.

Remark 1. If $m(\Pi, \Pi', \alpha) > i(\Pi, \Pi', \alpha)$, then we call Π' 'conservative', if $m(\Pi, \Pi', \alpha) < i(\Pi, \Pi', \alpha)$ then we call it 'liberal', and 'neutral' in case of equality. If $\alpha = 0$ (resp. $\alpha = 1$), then the evaluator prefers conservative rather than liberal (resp. liberal rather than conservative) adversarial partitions. Consider an adversary thinking that, by default, all elements are unlinked (resp. linked), and links (resp. unlinks) elements only if it observes strong evidence in support of this. Roughly speaking, in the presence of uncertainty, such an adversary is likely to make more miss

(resp. include) rather than include (resp. miss) errors, and thus end up with a conservative (resp. liberal) partition.

Remark 2. Note that $m(\Pi, \Pi', \alpha)$ and $i(\Pi, \Pi', \alpha)$ as well as their sum can be seen as distance measures between the two partitions. One should keep in mind, however, that they are asymmetric. If, for example, $\Pi = \{\{\circ, \bullet, \odot\}\}$ and $\Pi' = \{\{\circ, \bullet\}, \{\odot\}\}$, then $m(\Pi, \Pi', 1/2) = 1$ but $m(\Pi', \Pi, 1/2) = 0$, and $i(\Pi, \Pi', 1/2) = 0$ but $i(\Pi', \Pi, 1/2) = 3$. We do not consider this asymmetry to be a problem because we do not aim to specify a real distance metric over the partition space. It is nevertheless important that our measures behave consistently with real distance metrics; see section 5 for an examination in this respect.

Combining and normalising primary errors For a given policy parameter α and solution cluster C, the combined error count with respect to an adversarial partition Π' is defined by $e(C, \Pi', \alpha) = \alpha m(C, \Pi', \alpha) + (1 - \alpha) i(C, \Pi', \alpha)$. In order to normalise this combined error count, we consider the worst case, i.e. the adversarial partition that maximises it. Assuming that $n \geq 2$, the worst case occurs if $\alpha = 0$, $|C| = 1$ and Π' is a singleton: in this case, there occur $m(C, \Pi', \alpha) = |C| - 1 = 0$ miss errors and $i(C, \Pi) = n - |C| = n - 1$ include errors, and, thus, $n - 1$ errors in total. We therefore define the normalised error count of Π' with respect to C as $e^*(C, \Pi', \alpha) = {}^{e(C,\Pi',\alpha)}/_{n-1}$. Note that $0 \leq e^*(C, \Pi', \alpha) \leq 1$; it is zero if the adversary has made no errors that the evaluator cares about, and one in the worst case just described. Also note that normalisation is possible only if $n \geq 2$. The overall normalised error count is the average $e^*(\Pi, \Pi', \alpha) = \sum_{C \in \Pi} {}^{e^*(C,\Pi',\alpha)}/_{|\Pi|}$.

3.2 Secondary errors

If an adversarial partition contains many more clusters than the solution partition, then primary error counts are bound to give an incomplete picture, because they ignore the exact structure of all but the most relevant clusters. We are thus interested in more detailed evaluation, and, in particular, in the number of mergings that are required until a given percentage of elements of a particular solution cluster have been linked in the adversarial partition. The smaller the number of required mergings, the better the partition. However, every merging potentially brings with it foreign elements, and the sensitivity towards the presence of such elements is a matter of the evaluator's policy.

Moreover, we must decide in which order the mergings are performed until the target percentage is reached. Our `Plot` algorithm, shown in

Fig. 1, makes use of the evaluator's policy parameter in order to establish this ordering. Given a solution cluster C, the adversarial partition Π', the policy parameter α, and a bit b that indicates whether or not contamination is desirable from the subject's point of view (if $b = 1$ then contamination is desirable)[2], the algorithm produces a plot of two graphs that communicates the quality of the adversarial partition with respect to C in an intuitive way. The *linked* graph, in particular, shows how quickly consecutive mergings lead to element linking, and the *mixed* graph shows how quickly foreign elements are mixed into the cluster as a result of the same mergings. The worst-case complexity of our implementation of the **PlotForCluster** algorithm is $O(n^2)$.

PlotForCluster (input: C, Π', α, b)

1. For all $C' \in L(C, \Pi')$, compute the priority $p_\alpha(C, C')$. Then order $L(C, \Pi')$ according to the priority, highest value first. Resolve ties by giving priority to clusters with fewer foreign elements. Resolve remaining ties arbitrarily.
2. Start with an empty cluster X.
3. For values of $j = 1$ until $|L(C, \Pi')|$, do the following.
   (a) Merge X with the jth cluster from $L(C, \Pi')$.
   (b) Plot the data point $(j - 1, 1 - {}^{m(C,X)}/_{|C|})$ in the 'linked' graph.
   (c) If $|C| = n$, set $y = 0$. Otherwise set $y = {}^{i(C,X)}/_{n-|C|}$.
   (d) If $b = 0$ (resp. $b = 1$), plot the data point $(j - 1, y)$ (resp. $(j - 1, 1 - y)$) in the 'mixed' graph.

Fig. 1: Plotting the 'linked' and 'mixed' graphs for a given solution cluster

Based on these plots, we can measure how 'dangerously' the adversarial partition approaches any given threshold $\beta \in [0, 1]$, where β represents a percentage of to-be-linked elements of a given solution cluster C. This is done as follows. First, we let $y_{\lambda,C,\Pi'}(x)$ (resp. $y_{\mu,C,\Pi'}(x)$) denote the y-coordinate of the data point at x in that cluster's linked (resp. mixed) graph. If $\beta \leq y_{\lambda,C,\Pi'}(0)$, then the threshold β has already been surpassed by the partition Π'; no mergings are required to reach β. If $\beta > y_{\lambda,C,\Pi'}(0)$, on the other hand, then we proceed as follows. First, we draw a horizontal line starting at the point $(0, \beta)$. Using this line, we find the point $(x_{\lambda,C,\beta}, y_{\lambda,C,\beta})$ of the linked graph that corresponds to β. From there, we draw a vertical line and find the point $(x_{\mu,C,\beta}, y_{\mu,C,\beta})$ where this line meets the mixed graph. We call the vectors that start at the origin and point to $(x_{\lambda,C,\beta}, y_{\lambda,C,\beta})$ and $(x_{\mu,C,\beta}, y_{\mu,C,\beta})$, the '$\beta$-linked' and the '$\beta$-mixed' vectors, respectively. The *slope* of the β-linked vector, expressed

[2] The loan seeker (resp. consumer) from the introduction would set $b = 0$ (resp. $b = 1$).

as a percentage, is then used as a measure of how quickly the adversarial partition approaches the threshold.

Combining and normalising secondary errors Unless $\alpha = 0$, the slope of the β-linked vector must be co-evaluated with the slope of the β-mixed vector, because the latter expresses how quickly foreign elements 'contaminate' the adversarial cluster as it approaches the given threshold. We again use the policy parameter α in order to combine the two slopes into a single measurement: the risk slope of a given solution cluster C with respect to an adversarial partition Π', a threshold $\beta > y_{\lambda,C,\Pi'}(0)$, and policy parameter α is defined as $\Delta(C, \Pi', \beta, \alpha) = {}^2\!/\!_\pi[\alpha \arctan(y_{\lambda,C,\beta}/x_{\lambda,C,\beta}) + (1-\alpha)\arctan(y_{\mu,C,\beta}/x_{\mu,C,\beta})]$. Note that $0 \leq \Delta(C, \Pi', \beta, \alpha) \leq 1$.

Evaluating the entire partition The above computations and graphs measure the extent to which a given solution cluster is hidden within an adversarial partition. We would like to plot a single graph that summarises the situation for the entire solution partition and that somehow conveys the extent to which 'the typical' solution cluster is hidden in the solution partition. Unfortunately, taking the straight-forward average $|\Pi|^{-1} \sum_{C \in \Pi} \Delta(C, \Pi', \beta, \alpha)$ is not an option, because the values of $y_{\lambda,C,\Pi'}(0)$ are likely to differ for each C and this forces this expression to remain undefined for all $\beta \leq \max_{C \in \Pi} y_{\lambda,C,\Pi'}(0)$.

We circumvent this problem by plotting the 'overall' linked and mixed graphs using the **PlotForPartition** algorithm shown in Fig. 2. Note that, in order to make the algorithm work, the quantities $y_{\lambda,C,\Pi'}(x)$ and $y_{\mu,C,\Pi'}(x)$ for values of x between $|L(C, \Pi')|$ and $|\Pi| - 1$ (inclusive) must first be defined; recall that, since both graphs in the plot for cluster C have exactly $|L(C, \Pi')|$ data points (one representing the adversary's current state, and one for every merging until the adversary has linked all elements in C), these quantities were defined above only for $x < |L(C, \Pi')|$. We use the following recursive flat definitions to define $y_{\lambda,C,\Pi'}(x)$ and $y_{\mu,C,\Pi'}(x)$ for the missing range: for all $x \geq |L(C, \Pi')|$, $y_{\lambda,C,\Pi'}(x) \stackrel{\text{def}}{=} y_{\lambda,C,\Pi'}(|L(C, \Pi')| - 1) = 1$ and $y_{\mu,C,\Pi'}(x) \stackrel{\text{def}}{=} y_{\mu,C,\Pi'}(|L(C, \Pi')| - 1)$. If, for example, a plot does not contain a data point for the third merging because the adversary can fully link the elements of the cluster in, say, two mergings – $|L(C, \Pi')| = 3$ in this case – , then we take into account the situation after the final (i.e. second) merging. That is, the y-coordinates of all data points for three or more mergings, are defined to be identical to the y-coordinates of the data points at the second merging. For the linked graph, these coordinates are always 1 (since it is the *final* merging).

For the mixed graph, they are equal to the percentage of foreign elements that were present after the final merging.

Given this definition, the linked (resp. mixed) graph produced by **PlotForPartition** represents the average percentage of elements, over all subjects, that are linked (resp. mixed) after the adversary is given *an allowance* of performing up to x cluster mergings. Armed with these graphs, the risk slope $\Delta(\Pi, \Pi', \beta, \alpha)$ representing the 'average solution cluster' for a given threshold β can be computed in the same manner as $\Delta(C, \Pi', \beta, \alpha)$. The worst-case complexity of our implementation of **PlotForPartition** is $O(n^3)$

PlotForPartition (input: Π, Π', α)

1. Run **PlotForCluster** for all clusters $C \in \Pi$.
2. For values of $j = 0$ until $|\Pi| - 1$, do the following.
   (a) Compute the averages $\hat{y}_l = \frac{\sum_{C \in \Pi} y_{\lambda, C, \Pi'}(j)}{|\Pi|}$ and $\hat{y}_m = \frac{\sum_{C \in \Pi} y_{\mu, C, \Pi'}(j)}{|\Pi|}$.
   (b) Plot $(j, \hat{y}_l)$ and $(j, \hat{y}_m)$ in the 'linked' and 'mixed' graphs, respectively.

Fig. 2: Plotting the 'linked' and 'mixed' graphs for the entire partition

3.3 Sensitive elements

Merely counting errors presumes that all elements are equal. In reality, however, often only little harm is done if an adversary links some elements, as long as certain particularly sensitive ones remain unlinked. Similarly, one may be not suffer much if an adversary links some (or all) of one's elements, as long as certain foreign elements, perhaps of a particularly desirable type, are being mixed according to the adversary's view. In order to account for different element sensitivities, we enable the evaluator to first attach a weight $w_\ell \in [0, 1]$ to each element ℓ. The weights are required to represent the relative sensitivity of the elements. We therefore generalise the miss and error count formulas as $m(C, C') = \sum_{\ell \in \{C - C'\}} w_\ell$ and $i(C, C') = \sum_{\ell \in \{C' - C\}} w_\ell$; the remainder of our framework remains unchanged.

Note that subjects may disagree as to which elements are more sensitive than others. They may also disagree on the value that the policy parameter α should have. This is not a problem in our framework; the evaluator may assign both different sensitivities to the elements and different values to α and well as $b$for each solution cluster evaluation. That is, each cluster plot may have a different underlying sensitivities; since

step 2 of **PlotForPartition** does not take sensitivities into account, divergent sensitivities cause no problem. On the contrary, they will cause the overall plot to represent more accurately the summary of the risk as perceived by each subject.

3.4 Adversarial views over partitions

So far we have assumed that the adversary outputs a single partition. It may, however, output a *probability distribution* over the space of partitions. Note that a computationally bounded adversary can only output a distribution that can be encoded in polynomial length. Without loss of generality, we assume that an adversary outputs a view $\mathcal{V} = \{(\Pi'_1, \Pr(\Pi'_1 = \Pi)), (\Pi'_2, \Pr(\Pi'_2 = \Pi)), \dots\}$ such that , for all $1 \leq i \leq |\mathcal{V}|$, $\Pr(\Pi'_i = \Pi) > 0$ and $\sum_i \Pr(\Pi'_i = \Pi) = 1$. The pair $(\Pi'_i, \Pr(\Pi'_i = \Pi))$ means that, according to the adversary's view, Π'_i is the correct partition with probability $\Pr(\Pi'_i = \Pi)$.

Primary errors We define the average miss and include error counts for solution cluster C with respect to a view $\mathcal{V}$ as

$$m(C, \mathcal{V}, \alpha) = \sum_{((\Pi', \Pr(\Pi' = \Pi)) \in \mathcal{V}} \Pr(\Pi' = \Pi) m(C, \Pi', \alpha) \text{ and}$$

$$i(C, \mathcal{V}, \alpha) = \sum_{((\Pi', \Pr(\Pi' = \Pi)) \in \mathcal{V}} \Pr(\Pi' = \Pi) i(C, \Pi', \alpha),$$

respectively.[3] These formulas then replace $m(C, \Pi, \alpha)$ and $m(C, \Pi, \alpha)$, and the remainder of the primary error evaluation as described in section 3.1 remains unchanged.

Secondary errors In order to plot the linked and mixed graphs for a given solution cluster with respect to an adversarial view $\mathcal{V}$, we use the **PlotForClusterGivenView** algorithm shown in Fig. 3. This algorithm is very similar in spirit with the **PlotForPartition** algorithm; the difference is that the plotted values are not averages over clusters, but rather weighted averages over the partitions in the view. Note that, as expected, both **PlotForCluster** and **PlotForClusterGivenView** effectively yield identical plots for adversaries that output a single partition.

[3] These definitions follow the spirit of Equation 3 in [3], which also weighs a particular partition-dependent quantity by the probability that the underlying partition is the correct one.

Also note that the worst-case complexity of **PlotForClusterGivenView** is $O(n^2|\mathcal{V}|)$. That is, the evaluation of sufficiently lengthy adversarial views, i.e. such that $|\mathcal{V}| \gg n^2$, is roughly linear in their size.

PlotForClusterGivenView (input: $C, \mathcal{V}, \alpha$)

1. Run **PlotForCluster**(C, Π', α) for all $\Pi' \in \mathcal{V}$.
2. For values of $j = 0$ until $|\Pi'| - 1$, do the following.
   (a) Compute the weighted averages $\hat{y}_\lambda = \sum_{\Pi' \in \mathcal{V}} \Pr(\Pi' = \Pi) y_{\lambda, C, \Pi'}(j)$ and $\hat{y}_\mu = \sum_{\Pi' \in \mathcal{V}} \Pr(\Pi' = \Pi) y_{\mu, C, \Pi'}(j)$
   (b) Plot $(j, \hat{y}_l)$ and $(j, \hat{y}_m)$ in the 'linked' and 'mixed' graphs, respectively.

Fig. 3: Plotting the 'linked' and 'mixed' graphs of given solution cluster with respect to an adversarial view over the partition space

4 Example

Consider a set of 24 items and the solution partition Π that divides it into four clusters of equal size: triangles, squares, circles, and stars. Suppose that two adversaries, which are asked to partition the set, come up with the adversarial partitions shown in Fig. 4 (first row). Assuming that our policy parameter is $\alpha = 1/2$, the left partition Π'_1 is a conservative one, because it exhibits $m(\Pi, \Pi'_1, \alpha) = m(\text{triangles}, \Pi'_1, \alpha) + m(\text{squares}, \Pi'_1, \alpha) + m(\text{circles}, \Pi'_1, \alpha) + m(\text{stars}, \Pi'_1, \alpha) = 4 + 2 + 3 + 3 = 12$ miss errors, but only $i(\Pi, \Pi'_1, \alpha) = 4$ include errors. The right partition Π'_2 is a liberal one: it exhibits $m(\Pi, \Pi'_2, \alpha) = 7$ miss errors and $i(\Pi, \Pi'_2, \alpha) = 15$ include errors.

The star-star-circle cluster of Π'_1, for example, does not contribute to the primary error counts at all. We therefore evaluate the partitions with respect to secondary errors. The second row of Fig. 4 shows the output of **PlotForCluster** for the star cluster, and with respect to the two adversarial partitions Π'_1 (left) and Π'_2 (right). The left (resp. right) side plot also shows the 'linked' and 'mixed' vectors corresponding to the linkage of $\beta = 90\%$ (resp. $\beta = 70\%$) of the stars. We have that $\Delta(\Pi, \Pi'_1, 0.9, 1/2) \approx 1/2(0.364 + 0.100) \approx 23.20\%$ and $\Delta(\Pi, \Pi'_2, 0.7, 1/2) \approx 1/2(0.361 + 0.340) \approx 35.03\%$.

The third row of Fig. 4 shows the output of **PlotForPartition** with respect to the two adversarial partitions, for different values of the parameter α. Note that, for most parameter values, the linked and mixed graphs coincide partially or entirely. Finally, the fourth row of Fig. 4

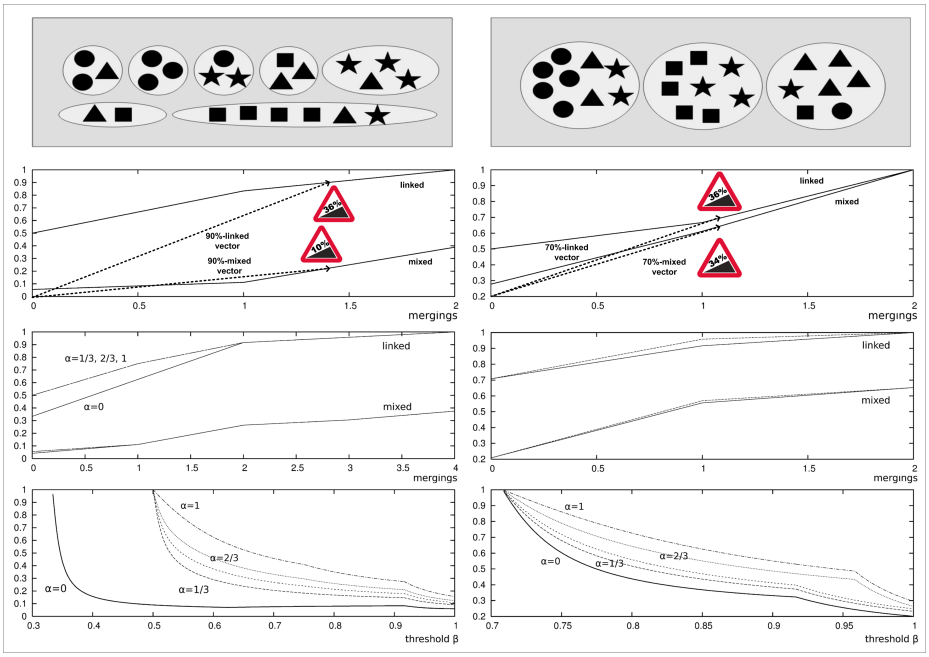

Fig. 4: First row: a conservative (left) and a liberal (right) adversarial partition. Second row: linked and mixed graphs for the star cluster and with respect to the adversarial partitions above ($\alpha = 1/2$). Third row: overall linked and mixed graphs, with respect to the adversarial partitions above, and for all $\alpha \in \{0, 1/3, 2/3, 1\}$. Fourth row: risk slopes $\Delta(\Pi, \Pi', \beta, \alpha)$ of the plots in the third row, as a function of the threshold β, for all $\alpha \in \{0, 1/3, 1/2, 2/3, 1\}$. Contamination is assumed to be undesirable (i.e. $b = 0$).

shows the the risk slopes, i.e. the value of $\Delta(\Pi, \Pi', \beta, \alpha)$ as a function of the threshold percentage β.

5 Simulated attacks

This section examines the behaviour of our risk slope measure. Our motivation is to demonstrate that it behaves intuitively when viewed as a distance measure between partitions. We first take a brief look at its behaviour in the setting of *uniformly at random* chosen partitions. (For an efficient way to choose partitions in this way, see chapter 10 of [10].) Due to space constraints we only show results for the case where contamination is undesirable, i.e. where $b = 0$. Figure 5 shows and constrasts how three distance measures behave for uniformly at random chosen partitions as the number of elements n grows. The three measures shown are $\Delta(\Pi, \Pi', \beta, \alpha)$ (for $\alpha = 1/3, 2/3$ and $\beta = 0.7, 0.8$), the variation of information (VOI) [7], and the minimal transfer distance (MTD) [2].

We observe that, similarly to the MTD and VOI, $\Delta(\Pi, \Pi', \beta, \alpha)$ behaves smoothly as n grows. This is important because a weak dependence on n is preferable to a strong one (see section 5 of [7]). Moreover, while the MTD and the VOI measures only depend on Π and Π', $\Delta(\Pi, \Pi', \beta, \alpha)$ varies depending on the sensitivities specified by the evaluator. The figure demonstrates that $\Delta(\Pi, \Pi', 0.7, \alpha) > \Delta(\Pi, \Pi', 0.8, \alpha)$. This matches our intuition that, since linking 70% of a subjet's elements is generally easier than linking 80%, the risk of this happening is higher, too. The figure also shows that $\Delta(\Pi, \Pi', \beta, 2/3) > \Delta(\Pi, \Pi', \beta, 1/3)$. This matches our intution that, since uniformly at random chosen partitions tend to be conservative (i.e. have many relatively small, rather than few very large clusters), attaching more importance to the presence of foreign elements results in lower risk levels. Finally, we observe that, as n grows, the MTD increases while the other measures decrease. The reasons for this lie as much with the measures themselves as with the nature of uniformly at random chosen partitions and the applied normalisations. Due to space constraints we do not analyse this further here. See appendix A for more information on the behaviour of our measures in the setting of random partitions.

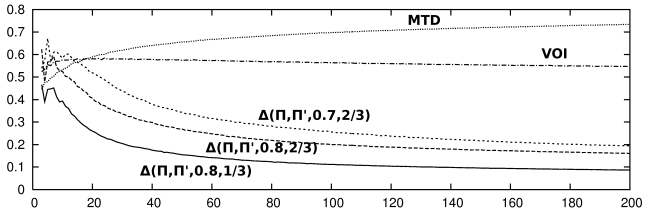

Fig. 5: $\Delta(\Pi, \Pi', \beta, \alpha)$, the variation of information normalised by $\log_2(n)$ (see section 5.1 of [7]), and the minimal transfer distance [2], normalised by $n - 1$. The shown results are averages of 1000 experiment repetitions.

In order to demonstrate that the behaviour of $\Delta(\Pi, \Pi', \beta, \alpha)$ is consistent and intuitive, we now compare it to the behaviour of the VOI distance in more detail. While we could use any reasonable measure as the basis for our comparison, we use the VOI because it has been shown to be a true metric [7]. We generate data for the comparison as follows. First, we choose a solution partition Π. Then we generate many partitions Π' that have different distances from Π. These partitions are generated by means of random walks of different lengths that start at Π and explore the solution space from there. Finally, we plot the $\Delta(\Pi, \Pi', \beta, \alpha)$ distance (vertical axis) against the VOI distance (horizontal axis) between Π and

each Π'. Since the partitions Π' have different distances from Π, this partition generation method simulates attacks on unlinkability of different effectiveness.

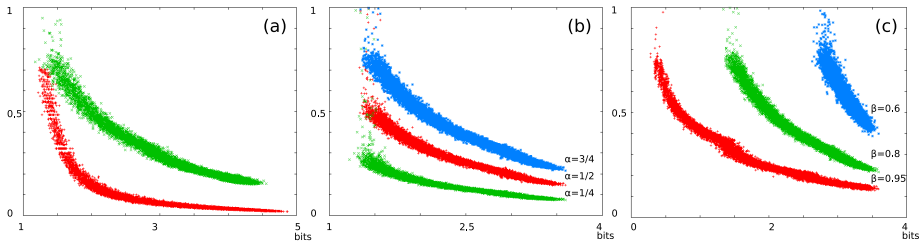

Fig. 6: The behaviour of $\Delta(\Pi, \Pi', \beta, \alpha)$ with respect to the VOI metric.

The three plots in Figure 6 show how $\Delta(\Pi, \Pi', \beta, \alpha)$ behaves as the VOI distance between partitions over a set of $n = 200$ elements increases. They demonstrate that the risk slope decreases monotonically as the VOI distance between partitions increases. This matches our intuition that, as the distance between the true partition and the adversary's guess increases, the average risk of reaching a particular threshold, also decreases.

Figure 6 (a) shows the effect of two selection methods for Π: the upper graph shows the simulation results when Π is chosen uniformly at random, and the lower graph shows the extreme case where Π is the singleton partition. For both graphs, parameters were set to $\alpha = 0.75$ and $\beta = 0.80$. Figure 6 (b) demonstrates the influence of the policy parameter α. The graphs show the cases for $\alpha = 1/4, 1/2$, and $3/4$, respectively. The underlying partition Π' was chosen uniformly at random and the threshold parameter was fixed at $\beta = 0.8$. Note that, as α decreases, $\Delta(\Pi, \Pi', \beta, \alpha)$ decreases. As explained above, this is due to the conservative nature of uniformly at random chosen partitions. Finally, Figure 6 (c) demonstrates the effect of β. The graphs show the cases for $\beta = 0.6, 0.8$, and 0.95, respectively. The underlying partition Π' was chosen uniformly at random and the policy parameter was fixed at $\alpha = 3/4$. The plot demonstrates the intuitive result that, as β increases, the risk slope decreases.

6 Concluding remarks

We introduced a framework for the evaluation of adversarial partitions. The framework is flexible because it enables the evaluator to attach different levels of importance to the adversary's inability to link a given subject's elements, and its inability to distinguish the subject's elements

from elements of other subjects. The evaluator may also specify, for each subject, different sensitivities for each element in the system as well as a threshold that represents what constitutes a privacy breach.

Our framework focuses on errors made by the adversary and distinguishes between primary and secondary errors. While primary errors measure how well a given subject's elements are currently linked and mixed with other subjects' elements, secondary errors project the risk that the adversary's current state represents for the future. Underlying this risk measurement is the implicit assumption that the adversary will obtain, in this future, information that enables a gradual merging of the clusters that contain some of that subject's elements, without at the same time filtering out foreign elements. In some settings, the adversary may be able to obtain information that leads to different ways of refining its current view. In such cases, secondary errors may turn out to be a less accurate representation of the real risk. We believe that, nevertheless, this does not invalidate the current approach because *any* refinement can be defined as a set of cluster mergings and splittings: our linked graph accounts for the mergings, and our mixed graph partially accounts for the splittings.

Our evaluation framework does not, however, fully account for the number of splittings necessary to divide all elements into clusters. This is because, while the mixed graph represents the number of foreign elements, it does not take into account the exact number of subjects that correspond to these foreign elements. In some settings, for example [6], this number, as well as the requirement that each other subject should be represented by an approximately equal number of foreign elements, is important. Refining the evaluation framework in this respect while retaining its intuitive and flexible nature, is future work.

We envision our measures to be used for the evaluation of attacks on unlinkability in diverse settings including anonymous communication, online anonymous transactions, and identity management. We expect them to be useful because, by enabling the evaluator to play with the α and β parameters on the subject level, they offer the ability to evaluate attacks in more detail. Of course, in order to visualise and evaluate the adversary's state, this state must first be gathered. While this is typically not a problem in experimental settings, doing this without turning the data gatherer himself into the adversary remains a challenge in most real-world settings. Sometimes, however, for example in the setting of database sanitisation, the adversary's state is published. We expect our measures, just like other privacy measures, to be useful in pre-deployment analysis and in cases where reliable information about the adversary's state is available.

Acknowledgements

The authors are grateful to Filipe Beato and Markulf Kohlweiss for their insightful comments on an earlier version of this paper. This paper describes work undertaken partly in the context of the 'Trusted Architecture for Securely Shared Services' (TAS3) project (www.tas3.eu). TAS3 is a collaborative project supported by the 7th European Framework Programme, with contract number 216287.

References

1. S. Clauß. A framework for quantification of linkability within a privacy-enhancing identity management system. In G. Müller, editor, *Emerging Trends in Information and Communication Security, International Conference, ETRICS 2006, Freiburg, Germany, June 6-9, 2006, Proceedings*, volume 3995 of *Lecture Notes in Computer Science*, pages 191–205. Springer Verlag, 2006.
2. L. Denœud and A. Guénoche. Comparison of distance indices between partitions. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, editors, *Data Science and Classification*, Mathematics and Statistics, pages 21–28. Springer Verlag, Berlin, 2006.
3. L. Fischer, S. Katzenbeisser, and C. Eckert. Measuring unlinkability revisited. In *Proceedings of the 2008 ACM Workshop on Privacy in the Electronic Society, WPES 2008, Alexandria, Virginia, USA, October 27, 2008*, pages 111–116. ACM Press, 2008.
4. M. Franz, B. Meyer, and A. Pashalidis. Attacking unlinkability: The importance of context. In N. Borisov and P. Golle, editors, *Privacy Enhancing Technologies, 7th International Symposium, PET 2007 Ottawa, Canada, June 20-22, 2007, Revised Selected Papers*, volume 4776 of *Lecture Notes in Computer Science*, pages 1–16. Springer Verlag, Berlin, 2007.
5. J. Kogan. *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, 2007.
6. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.
7. M. Meilă. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895, 2007.
8. M. Meilă. Comparing clusterings: an axiomatic view. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 577–584, New York, NY, USA, 2005. ACM Press.
9. M. Neubauer. Modelling of pseudonymity under probabilistic linkability attacks. *IEEE International Conference on Computational Science and Engineering*, 3:160–167, 2009.
10. A. Nijenhuis and H. S. Wilf. *Combinatorial Algorithms*. Academic Press Inc, 2nd edition, 1978.
11. A. Pashalidis. Measuring the effectiveness and the fairness of relation hiding systems. In *Proceedings of the First International Workshop on Multimedia, Information Privacy and Intelligent Computing Systems*, 2008.

12. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
13. S. Schiffner and S. Clauß. Using linkability information to attack mix-based anonymity services. In *Privacy Enhancing Technologies, 9th International Symposium, PETS 2009, Seattle, WA, USA, August 5-7, 2009. Proceedings.*
14. S. Steinbrecher and S. Köpsell. Modelling unlinkability. In R. Dingledine, editor, *Privacy Enhancing Technologies, Third International Workshop, PET 2003, Dresden, Germany, March 26-28, 2003, Revised Papers*, volume 2760 of *Lecture Notes in Computer Science*, pages 32–47. Springer Verlag, Berlin, 2003.
15. D. Wagner and S. Wagner. Comparing clusterings–an overview. Technical Report 2006-04, Faculty of Informatics, University of Karlsruhe (TH), 2006.

A Random partitions

This section experimentally examines how our measures behave for randomly chosen adversarial partitions. This examination provides reference points to attack evaluators which can be used to quantify by how much an adversary outperforms random guessing. We perform two experiments which we call the 'uniform' and the 'non-uniform' experiment, respectively. In both experiments, Π and Π' are randomly generated from the space of all set partitions of sets of size n. The experiments differ in the way the partitions are drawn from the space. In the uniform experiment, in particular, Π and Π' are chosen *uniformly* from the space of partitions. In the non-uniform experiment, Π and Π' are generated as follows. Initially, the to-be-generated partition consists of a 'cluster population' containing a single cluster that contains the first element. The remaining $n-1$ elements are then assigned to clusters, one by one, as follows. For each element, a fair coin is tossed. In case of heads, a new cluster is created, the element is assigned to that cluster, and the cluster is added to the cluster population; otherwise, a cluster already in the population is chosen uniformly at random and the element is assigned to it.

Fig. 7 shows some experiment results. As far as primary errors are concerned, in the uniform experiment occur, on average, slightly more errors than in the non-uniform experiment. Observe that, unless α takes very high values, in both experiments there occur less include than miss errors. For high values for α (say, above 90%), there occur more include rather than miss errors. This is because foreign elements are largely disregarded when determining the most relevant cluster and, as a result, that cluster has more foreign than correct elements. The average combined normalised error counts reflect the fact that there occur slightly fewer primary errors in the non-uniform than in the uniform experiment. Observe that, in our experiments, the choice of α has no significant impact on this measure.

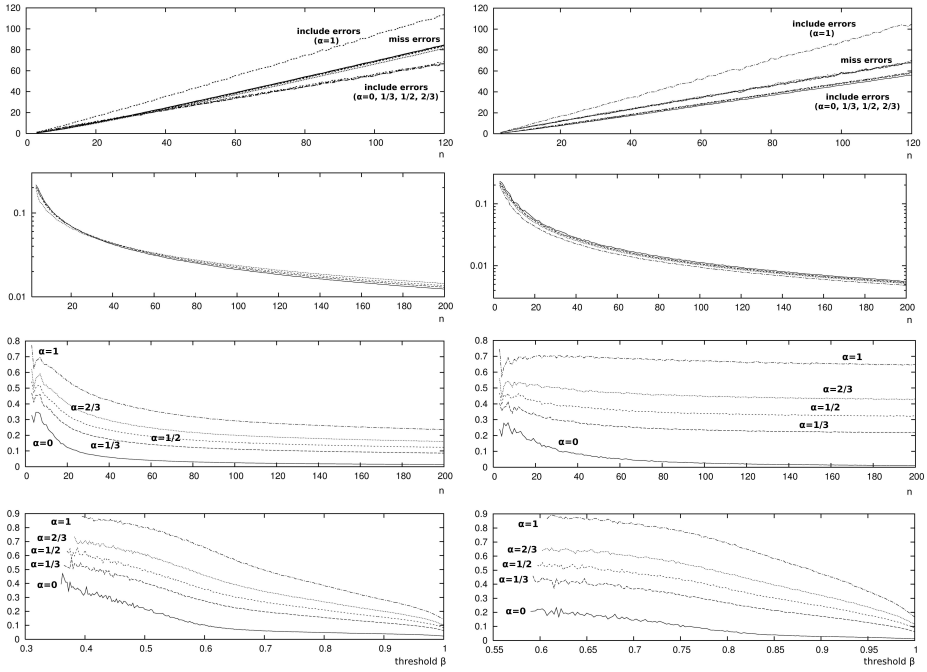

Fig. 7: First row: Average number of miss and include errors, i.e. $m(\Pi, \Pi', \alpha)$ and $i(\Pi, \Pi', \alpha)$, for varying n. Second row: normalised combined error counts $e^*(\Pi, \Pi', \alpha)$ on a logarithmic scale, for varying n. Third row: Risk slope $\Delta(\Pi, \Pi', 0.8, \alpha)$ for varying n. Fourth row: Risk slope $\Delta(\Pi, \Pi', \beta, \alpha)$, for constant $n = 50$ and varying threshold β. The results shown are for all parameter values $\alpha \in \{0, 1/3, 1/2, 2/3, 1\}$ and contamination is assumed to be undesirable ($b = 0$). Plots on the left and right hand side show results from the uniform and non-uniform experiment, respectively.

The value of α has, on the other hand, a significant impact on the risk slopes $\Delta(\Pi, \Pi', \beta, \alpha)$; higher values of α yield higher risk slopes. Moreover, in the uniform experiment, the difference between the risk slopes for low and high values for α is much smaller than the corresponding difference in the non-uniform experiment. Finally, as the threshold β increases, the risk slopes converge. In the uniform experiment they converge slowly; in the non-uniform experiment they do not converge until $\beta \approx 0.7$, but then they converge fast.